Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 49 (2005) 217–226
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Equivalence of two treatments and sample size determination under exponential survival model with censoring

Jun-mo Nam[a], Jinheum Kim[b], Seungyeoun Lee[c,*]

[a]*Department of Health and Human Services, Biostatistics Branch, DCEG, National Cancer Institute, National Institutes of Health, Rockville, MD 20852-7240, USA*
[b]*Department of Applied Statistics, University of Suwon, Gyeonggi-Do 445-743, Republic of Korea*
[c]*Department of Applied Mathematics, Sejong University, 98 Gunja-dong, Gwangjin-gu, Seoul, 143-747, Republic of Korea*

Received 9 February 2004; received in revised form 30 April 2004

Abstract

We present the likelihood score and F statistics for ascertaining equivalence of two treatments in survival under an exponential model with independent censoring. We provide explicit formulae for power and sample size requirement for trials using the score and F tests, and compare the score and F tests with the log rank test by Com-Nougue et al. (Statist. Med. 12 (1993) 1353). Simulation results show that empirical powers of the score, F and log rank tests are satisfactorily close to the corresponding asymptotic powers for small-to-moderate sample size. We find these three methods are essentially identical in terms of level and power. However, the score and F methods are very sensitive to departure from the exponential assumption while the log rank test is more robust. The methods are illustrated by application to data from a randomized trial of two treatments for B non-Hodgkin lymphoma. Published by Elsevier B.V.

Keywords: Censoring; Equivalence trial; Exponential model; Score test; F test; Power and sample size

1. Introduction

Recently, there has been a great interest in establishing equivalence of two treatments in clinical trials. For example, a standard chemotherapy in pediatric oncology is highly

* Corresponding author. Tel.: +82-2-3408-3161; fax: +82-2-3408-3315.
E-mail addresses: namj@mail.nih.gov (J.-m. Nam), jinhkim@suwon.ac.kr (J. Kim), leesy@sejong.ac.kr (S. Lee).

0167-9473/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.csda.2004.05.009

effective but causes severe toxic side-effects, and researchers are interested in a less toxic new treatment which may be essentially as effective as the standard one in survival (Patte et al., 1991). The intention of an equivalence trial is to demonstrate that two treatments do not differ by more than a prescribed small amount which is materially insignificant. The conventional test procedure for detecting a difference in a comparative trial cannot be applied for this situation.

Statistical methods for establishing one-sided equivalence or non-inferiority of a new treatment to the standard one on binary responses have been investigated by many authors, e.g., Dunnett and Gent (1977), Roebruck and Kühn (1995) and Nam (1997). For equivalence of two survival distributions with censored observations, Wellek (1993) and Com-Nougue et al. (1993) have proposed testing procedures based on the proportional hazards model. Wellek (1993) derived the uniformly most powerful test in terms of the maximum partial likelihood estimator but the sample size equation is not given explicitly while Com-Nougue et al. (1993) provided the confidence intervals for the actual hazard ratio based on the log rank test statistic. When data follow an exponential model with no censoring, Bristol and Desu (1990) have suggested a parametric method of testing for equivalence. However, a parametric method based on censored data has not been thoroughly studied.

In this paper, we investigate statistical methods involving the equivalence of two treatments based on exponentially distributed survival data with censoring. In Section 2, we derive two different tests for equivalence: the score test and F test procedures. In addition, the asymptotic powers and approximate sample size formula are provided. In Section 3, the score and F tests are compared with the log rank test (Com-Nougue et al., 1993) by simulations in level and power, and approximate numbers of events required for a specific power using these methods are examined. Also, we investigate the robustness of the three tests when the underlying exponential model is violated. Sections 4 and 5 contain an example based on non-Hodgkin's malignant type B lymphoma data and discussion.

2. Test statistics and power functions

2.1. Score method

Denote the survival and censoring times of standard and new treatment groups by t_{ij} and c_{ij} for $i = 0, 1$ and $j = 1, 2, \ldots, n_i$, respectively. The first subscript $i = 0, 1$ indicates the standard and new treatment groups and the second subscript indicates the jth individual in the ith group. Under right censorship, we observe survival data $\{(x_{ij}, \delta_{ij}), i = 0, 1$ and $j = 1, 2, \ldots, n_i\}$, where $x_{ij} = \min(t_{ij}, c_{ij})$ and $\delta_{ij} = I(t_{ij} \leqslant c_{ij})$, i.e., $\delta_{ij} = 1$ if $t_{ij} \leqslant c_{ij}$ and $\delta_{ij} = 0$ otherwise. Assume that t_{ij} and c_{ij} are independent within a group.

Consider the exponential survival distributions of the standard and the new treatment groups as $S_0(t) = \exp(-h_0 t)$ and $S_1(t) = \exp(-h_1 t)$, where $h_i > 0$ for $i = 0, 1$. Denote the hazard ratio by $r = h_1/h_0$. Let $x_{i\cdot} = \sum_{j=1}^{n_i} x_{ij}$ and d_i be total survival follow-up time and the number of uncensored observations for $i = 0, 1$, respectively. The score statistic for testing $H_0 : r \geqslant r_0$ against $H_1 : r < r_0$ can be simplified as

$$z = \{d_1 - \widehat{\mu}(r_0)\}/\{\widehat{v}(r_0)\}^{1/2}, \tag{1}$$

where $\widehat{\mu}(r_0) = (d_0 + d_1)\{r_0 x_{1\cdot}/(x_{0\cdot} + r_0 x_{1\cdot})\}$ and $\widehat{v}(r_0) = d_0 d_1/(d_0 + d_1)$. We reject H_0 against H_1 at the significance level of α when $z \leqslant -z_{(1-\alpha)}$, where $z_{(1-\alpha)}$ is the $100(1-\alpha)$ percentile point of the standard normal distribution.

The asymptotic power of the score statistic (1) for testing $r = r_0$ against $r = r_1 (< r_0)$ is

$$Pr\{z \leqslant -z_{(1-\alpha)} | H_1 : r = r_1\} = 1 - \Phi(u), \tag{2}$$

where $u = [v(r_0)^{1/2} z_{(1-\alpha)} - \{\mu(r_1) - \mu(r_0)\}]/v(r_1)^{1/2}$. Denoting the design parameter by $\phi = n_1\pi_1/(n_0\pi_0)$ and given that the number of total events, d, is kept fixed, we have the expectation of d_1 under the null and the alternative as $\mu(r_0) = d\phi/(\phi + 1)$ and $\mu(r_1) = dr_0\phi/(r_0\phi + r_1)$ and the variances of d_1 under H_0 and H_1 as $v(r_0) = d\phi/(\phi + 1)^2$ and $v(r_1) = dr_0 r_1\phi/(r_0\phi + r_1)^2$, respectively. The approximate total number of uncensored observations required for a power, $1 - \beta$, of the score test at α is obtained by solving the following general equation with respect to d:

$$v(r_0)^{1/2} z_{(1-\alpha)} + v(r_1)^{1/2} z_{(1-\beta)} = \mu(r_1) - \mu(r_0), \tag{3}$$

e.g., Nam (1987). For a balanced design with equal censoring patterns for the standard and new treatment groups, formula (3) leads to

$$d = \{(r_0 + r_1) z_{(1-\alpha)} + 2(r_0 r_1)^{1/2} z_{(1-\beta)}\}^2/(r_0 - r_1)^2. \tag{4}$$

2.2. F method

For a general censoring pattern, the exact distribution of $u_i = 2n_i h_i/\widehat{h}_i$ where $\widehat{h}_i = n_i/x_{i\cdot}$ is unknown. However, we may have a good approximation by considering $u_i = 2d_i h_i/\widehat{h}'_i$, where $\widehat{h}'_i = d_i/x_{i\cdot}$ as a chi-square with $2d_i$ degrees of freedom, e.g., Cox and Oakes (1984). The ratio of two chi-squares standardized by their degrees of freedom, i.e., $F(r) = r\bar{x}_{1\cdot}/\bar{x}_{0\cdot}$, where $\bar{x}_{i\cdot} = x_{i\cdot}/d_i$ for $i = 0, 1$, is distributed approximately F with $2d_1$ and $2d_0$ degrees of freedom. We reject H_0 in a favor of H_1 at level α if

$$F(r_0)^{-1} = (r_0\bar{x}_{1\cdot}/\bar{x}_{0\cdot})^{-1} \leqslant F_{2d_0,2d_1}(\alpha), \tag{5}$$

where $F_{2d_0,2d_1}(\alpha)$ is the 100α percentile point of the F distribution with $2d_0$ and $2d_1$ degrees of freedom.

The asymptotic power of the F test for $r = r_0$ against $r = r_1 (< r_0)$ for given values of d_0 and d_1 at level α is

$$Pr\{F(r_0)^{-1} \leqslant F_{2d_0,2d_1}(\alpha) | H_1 : r = r_1\} = 1 - \beta. \tag{6}$$

A pair of event sizes with d_0 and d_1 may be found by (6) for a given power but the solution is not unique. Using a design parameter ϕ, we may approximately write $d_1 = \phi d_0$, where ϕ is fixed and power (6) is a function of d_0. In a typical case of $\phi = 1$, d_0 is obtained by the relation of

$$Pr\{F_{2d_0,2d_0} \leqslant (r_0/r_1) F_{2d_0,2d_0}(\alpha)\} = 1 - \beta. \tag{7}$$

3. Numerical evaluation

3.1. Size and power of tests

In order to compare the score and F tests with a non-parametric method, we introduce the log rank test derived by Com-Nougue et al. (1993). Denote $t_1, t_2, \ldots, t_d$ as the distinct times of events that occurred in the pooled treatment groups. For testing the equivalence of two treatments using survival data with censored observations, Com-Nougue et al. (1993) presented the modified log rank test statistic given by

$$u(r_0) = \{d_1 - E_1(r_0)\}/\{\mathrm{var}_1(r_0)\}^{1/2}, \tag{8}$$

which is asymptotically standard normal. Here $E_1(r_0) = \sum_{k=1}^{d} n_{1k} r_0/(n_{1k} r_0 + n_{0k})$ and $\mathrm{var}_1(r_0) = \sum_{k=1}^{d} n_{0k} n_{1k} r_0/(n_{1k} r_0 + n_{0k})^2$, where n_{ik} is the number of subjects at risk in the ith group at time t_k for $i = 0, 1$ and $k = 1, 2, \ldots, d$. For testing $\mathrm{H}_0 : r \geqslant r_0$ against $\mathrm{H}_1 : r < r_0$, we reject H_0 when $u(r_0)$ is small. From the result of Fleming (1990), an approximate number of events required for power $1 - \beta$ is derived as

$$d = 4\{z_{(1-\alpha)} + z_{(1-\beta)}\}^2/\{\ln(r_0) - \ln(r_1)\}^2. \tag{9}$$

Let $\Psi_k = n_{1k}/n_{0k}$ for $k = 1, 2, \ldots, d$. When $\Psi_k = 1$ for all k, (9) reduces to

$$d = \{r_0^{1/2}(r_1 + 1)z_{(1-\alpha)} + r_1^{1/2}(r_0 + 1)z_{(1-\beta)}\}^2/(r_1 - r_0)^2. \tag{10}$$

We conducted a simulation study to evaluate the performance of the score, F and log rank tests for equivalence of two treatments using survival data with independent censoring under a balanced design when the sample size is small or moderate. We generate the survival times of the standard and new treatment groups from exponential distributions with mean 1 and $1/r$, respectively, and also the censoring times from exponential distributions with means corresponding to a common censoring fraction. By matching survival and censoring times in n pairs in sequence, we generate the observed times, $x_{ij} = \min(t_{ij}, c_{ij})$, for $i = 0, 1$ and $j = 1, 2, \ldots, n_i$. Table 1 summarizes results of the computation of empirical size and power of the three tests based on 10,000 simulations for various values of the null and alternative, censoring fraction of 0.1, 0.3 and sample size $n = 25$ and 50. It shows that the empirical size and power are generally very close to the 5% level and the asymptotic power for each of the three tests. The power of a test is positively related to the difference between null and alternative and the sample size, and inversely related to the censoring fraction. The three tests are generally comparable in level and power.

From (4), (7) and (10), we calculate the approximate number of events required for power $= 60\%$ and 80% of the score, F and log rank tests at 0.05 level for various values of null and alternative in a balanced equivalence trial and present results in Table 2 . In addition, we simulate the actual power of a test for the required number of events given by the formula and add in Table 2. The three methods provide a similar requirement for total number of events except the cases of $r_0 = 2.0$ and $r_1 = 1.8$ and 1.5. In particular, those by the score and F methods produce virtually identical results. As the alternative approaches closer to the null, the number of events required for a specific power of the log rank test is slightly larger than required for the parametric methods. Fleming's method (1990), (9), also has the

Table 1
Empirical sizes and powers of score (S), F and log rank (ln R) tests based on 10,000 simulations for the sample size of $n_0 = n_1 = 25, 50$ when the underlying distribution is exponential

$n_0 = n_1$	r_0	r_1	Censoring fraction					
			0.1			0.3		
			S	F	ln R	S	F	ln R
25	2.0	2.0	0.054	0.050	0.051	0.054	0.052	0.052
			(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
		1.5	0.251	0.242	0.240	0.223	0.214	0.209
			(0.244)	(0.244)	(0.251)	(0.210)	(0.209)	(0.218)
		1.0	0.761	0.749	0.737	0.655	0.640	0.635
			(0.735)	(0.745)	(0.753)	(0.636)	(0.648)	(0.663)
	1.5	1.5	0.053	0.049	0.051	0.052	0.050	0.050
			(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
		1.0	0.392	0.381	0.385	0.330	0.317	0.321
			(0.378)	(0.382)	(0.394)	(0.319)	(0.321)	(0.334)
		0.8	0.683	0.671	0.670	0.582	0.566	0.573
			(0.661)	(0.671)	(0.684)	(0.565)	(0.576)	(0.591)
50	2.0	2.0	0.054	0.052	0.052	0.047	0.046	0.044
			(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
		1.5	0.398	0.391	0.375	0.333	0.327	0.318
			(0.385)	(0.387)	(0.385)	(0.325)	(0.326)	(0.328)
		1.0	0.947	0.945	0.937	0.898	0.892	0.888
			(0.946)	(0.948)	(0.946)	(0.888)	(0.892)	(0.892)
	1.5	1.5	0.051	0.050	0.051	0.051	0.051	0.050
			(0.050)	(0.050)	(0.050)	(0.050)	(0.050)	(0.050)
		1.0	0.611	0.604	0.599	0.531	0.523	0.523
			(0.602)	(0.606)	(0.612)	(0.512)	(0.516)	(0.525)
		0.8	0.907	0.904	0.902	0.834	0.827	0.825
			(0.904)	(0.907)	(0.911)	(0.828)	(0.834)	(0.841)

The values in parentheses are large sample power approximation with a nominal 0.05 level.

similar result as those of the score and F methods. The actual power of the score test for the approximate number of events given by (4) is greater than or equal to a nominal power. The number of events given by (7) and (10) provide the actual powers of F and log rank tests which are smaller than the nominal power except in a few cases. Fleming's method always yields an actual power that is less than the nominal one.

3.2. Robustness

The Weibull distribution with scale and shape parameters, ρ and κ, is flexible and has a wide application in survival data. Its survival distribution, density function and cumulative hazard rate are $S(t) = \exp\{-(\rho t)^{\kappa}\}$, $f(t) = \kappa\rho(\rho t)^{\kappa-1}\exp\{-(\rho t)^{\kappa}\}$, and $H(t) = (\rho t)^{\kappa}$. The hazard rate of Weibull is monotonically increasing, constant or decreasing according to $\kappa > 1$, $\kappa = 1$, and $\kappa < 1$. The exponential distribution is a special case of Weibull, i.e., $\kappa = 1$. We generate the survival times of standard and new treatment groups from Weibull

Table 2
Total number of events in trials calculated by the score (S), F, log rank ($\ln R$) and Fleming methods for power $= 60\%$, 80% and $\alpha = 0.05$

r_0	r_1	Nominal power (%)	S	F	$\ln R$	Flemming
2.0	1.8	60	1300	1300	1419	1298
			(0.605)	(0.600)	(0.599)	(0.567)
		80	2230	2229	2451	2228
			(0.803)	(0.795)	(0.811)	(0.757)
	1.5	60	176	175	182	174
			(0.600)	(0.599)	(0.592)	(0.574)
		80	301	300	317	299
			(0.806)	(0.806)	(0.795)	(0.783)
	1.2	60	57	56	55	55
			(0.610)	(0.595)	(0.574)	(0.578)
		80	97	96	97	95
			(0.805)	(0.794)	(0.782)	(0.778)
	1.0	60	32	31	29	30
			(0.621)	(0.596)	(0.558)	(0.577)
		80	54	52	52	51
			(0.802)	(0.789)	(0.783)	(0.774)
	0.8	60	19	16	16	17
			(0.629)	(0.530)	(0.550)	(0.576)
		80	31	30	29	29
			(0.812)	(0.786)	(0.756)	(0.757)
1.5	1.2	60	291	290	292	289
			(0.602)	(0.597)	(0.593)	(0.594)
		80	499	497	504	497
			(0.801)	(0.801)	(0.798)	(0.795)
	1.0	60	90	89	87	88
			(0.607)	(0.583)	(0.583)	(0.590)
		80	152	151	150	150
			(0.800)	(0.790)	(0.796)	(0.795)
	0.8	60	38	37	36	36
			(0.617)	(0.596)	(0.572)	(0.569)
		80	65	64	62	63
			(0.811)	(0.786)	(0.782)	(0.795)

The numbers in parentheses are the actual power of tests for the total number of events when censoring fraction =30%.

distributions with parameters, $(1, \kappa)$ and (ρ, κ), respectively, along with the corresponding censoring times. We examine actual levels of significance of the score, F and log rank tests from (1), (5) and (8), respectively, when the underlying model is Weibull. Simulation results are summarized in Table 3 . It shows that the actual levels of the score and F tests are anti-conservative when a hazard rate is accelerating ($\kappa > 1$) or they are conservative when a hazard rate is decelerating ($\kappa < 1$). If a hazard rate is nearly constant, the actual level is reasonably close to a nominal 0.05. The empirical level of the log rank test is satisfactorily close to 0.05. The reason is that the Weibull distribution accommodates the assumption

Table 3
Empirical sizes of score (S), F and log rank ($\ln R$) tests based on 10,000 simulations for the sample size of $n_0 = n_1 = 25, 50$ when the underlying distribution is Weibull with scale and shape parameters, (ρ, κ)

$n_0 = n_1$	(ρ, κ)	Censoring fraction					
		0.1			0.3		
		S	F	$\ln R$	S	F	$\ln R$
25	(2.46, 1/1.3)	0.032	0.033	0.051	0.027	0.027	0.049
	(2.30, 1/1.2)	0.035	0.036	0.048	0.035	0.035	0.048
	(2.14, 1/1.1)	0.042	0.043	0.052	0.043	0.043	0.052
	(2, 1)	0.050	0.050	0.052	0.052	0.052	0.054
	(1.88, 1.1)	0.061	0.063	0.058	0.062	0.062	0.054
	(1.78, 1.2)	0.069	0.071	0.052	0.067	0.068	0.049
	(1.70, 1.3)	0.081	0.084	0.053	0.082	0.083	0.055
50	(2.46, 1/1.3)	0.018	0.019	0.049	0.018	0.018	0.052
	(2.30, 1/1.2)	0.024	0.025	0.049	0.025	0.025	0.049
	(2.14, 1/1.1)	0.033	0.034	0.046	0.034	0.034	0.049
	(2, 1)	0.051	0.051	0.050	0.051	0.051	0.050
	(1.88, 1.1)	0.066	0.068	0.047	0.069	0.070	0.049
	(1.78, 1.2)	0.096	0.098	0.052	0.093	0.094	0.052
	(1.70, 1.3)	0.126	0.128	0.050	0.118	0.119	0.048

of the log rank test, i.e., a constant rates of the two hazard rates with respect to survival time.

4. An example

Consider an equivalence trial of two treatments for non-Hodgkin's malignant type B lymphoma (Com-Nougue et al., 1993). After an induction treatment of 4 months, the standard procedure provides a maintenance chemotherapy of 7 months while the new one shortens the length of the maintenance treatment to 4 months in order to reduce toxicity. Researchers would like to show that event-free survival rate is not compromised by the shortening the maintenance chemotherapy. The 166 patients were randomly assigned to treatments: 82 for the short arm and 84 for the long arm. The number of events that occurred in the new and standard treatment groups were $d_1 = 9$ and $d_0 = 11$. A patient having complete remission and no relapse within 18 months is considered cured, and there is no loss to follow-up before 18 months in the updated lymphoma data. Kaplan–Meier survival curves (1958) for the short and long arms are shown in Fig. 1. We may consider the short arm as effective as the long arm if the ratio of hazard rates (the former to the latter) is less than, say, 2. One-sided score and F tests for $H_0 : r \geqslant 2$ against $H_1 : r < 2$ yield $z = -1.989 (p = 0.023)$ and $F_{22,18} = 0.4006 (p = 0.022)$ from (1) and (5), and the log rank method gives $u = -1.981 (p = 0.024)$ from (8). Results of the three tests are essentially identical and reject $r \geqslant 2$ in a favor of $r < 2$. For $H_0 : r \geqslant 2.73$ against $H_1 : r < 2.73$ considered by Com-Nougue et al. (1993), the score, F and log rank tests yield highly significant p-values, i.e., 0.0051, 0.0037 and 0.0028, respectively. Significance testing indicates that

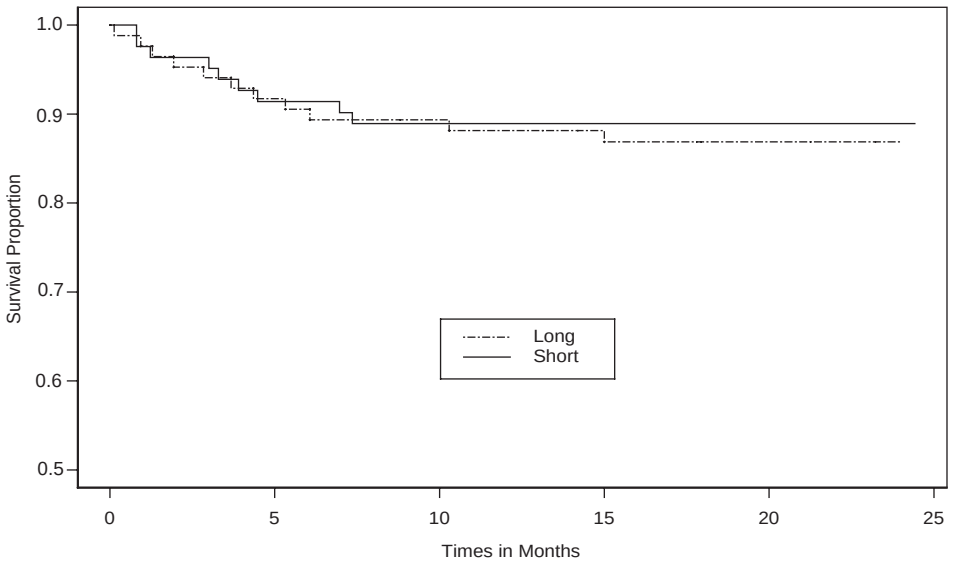

Fig. 1. Kaplan–Meier's survival curves of short(4 months) and long(7 months) maintenance treatments.

the short arm is not inferior to the long arm. In a previous trial by Patte et al. (1991), the 18 months survival rate of patients treated by the standard procedure was estimated as 90%. For a 90% event-free survival rate for the long arm, the relative risks of $r_0 = 2$ and 2.73 correspond to 9% and 15% lower survival rates for the short arm under an exponential model.

The estimated relative risk under the exponential model is $\widehat{r} = 0.8$ and the ratio of events of the new treatment group to those of the standard one is $\phi = d_1/d_0 = 0.818$. The powers of the score and F tests for testing $r_0 = 2$ against $r_1 = 0.8$ at $\alpha = 0.05$ are both 64.4% from (2) and (6), and those for $r_0 = 2.73$ against $r_1 = 0.8$ are 85.3% and 85.1%, respectively. The corresponding values of the power of the log rank test are 66.5% and 87.4%. The study has reasonably good power.

As shown in Fig. 2, the cumulative hazard rates of both treatment groups were approximately linear when $t \leqslant 7.5$ months. Since 90% of events occurred in this period, the exponential model is not inconsistent with the data. After 7.5 months, no relapse was observed for the short arm but two events occurred for the long arm. This suggest that longer exposure to toxicity may be rather detrimental.

5. Discussion

We investigated statistical methods for establishing the equivalence of two treatments based on exponentially distributed data with censoring. Asymptotic and empirical results show that the score and F tests are essentially the same in terms of level of significance and power. We found that the log rank test is also similar to the score and F tests for

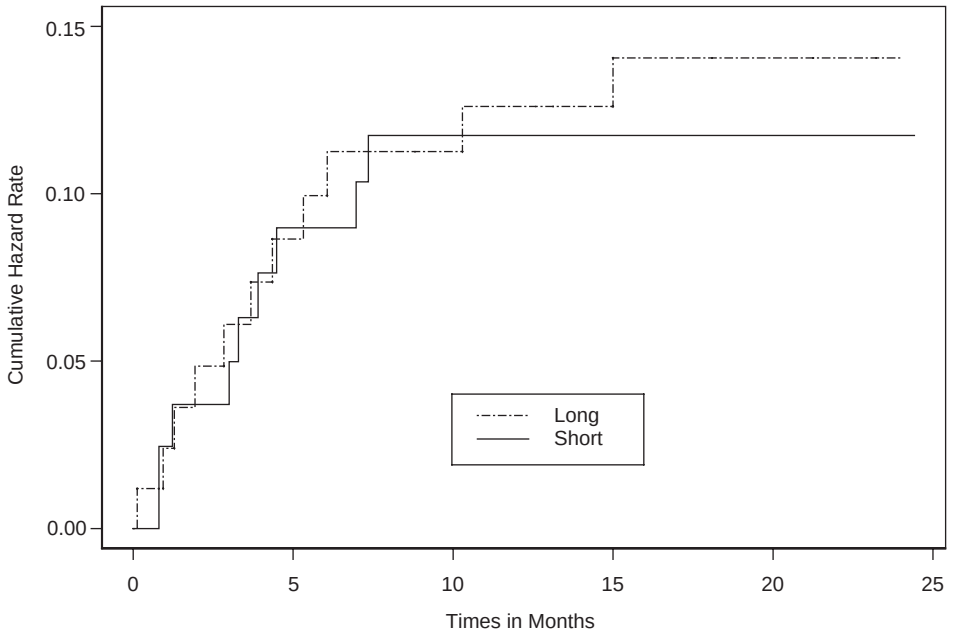

Fig. 2. Cumulative hazard rates of short(4 months) and long(7 months) maintenance treatments.

equivalence. We limited our simulations and numerical evaluation to equal censoring and a balanced design. Note that a randomized trial usually results in an allocation of an equal number of patients to the new and standard treatment groups.

The parametric tests are derived under the exponential model and the non-parametric method is obtained by assuming a constant ratio of two hazard rates over time. It is advisable to examine the adequacy of the model for the observed data by plotting the cumulative hazard rate over time. It is well known that the exponential distribution has been popular in the field of industrial reliability. Although the exponential model has found limited use in biomedical applications, it has been commonly employed for short term survival analysis where an aging process can be considered as a constant.

The two-parameter Weibull distribution is more flexible than the one-parameter exponential model in fitting survival data. Under the Weibull model, the score and F procedures for equivalence of two treatments based on the exponential model are biased except a special case (i.e., a constant hazard rate) while the log rank test is unbiased for all cases. The three tests can be applied for a short term survival analysis but we caution use of the parametric methods for a long term survival analysis.

Acknowledgements

The authors are grateful to Dr. C. Rodary of Institute Gustave Roussy in France for providing detailed survival data from a randomized trial of treatments for B non-Hodgkin

lymphoma. We thank an Associate Editor and referees for comments which helped to improve the presentation of this paper. This work was supported by a Korea Research Foundation Grant (KRF-99-042-D00025 D1208).

References

Bristol, D.R., Desu, M.M., 1990. Comparison of two exponential distributions. Biometrical J. 32, 267–276.

Com-Nougue, C., Rodary, C., Patte, C., 1993. How to establish equivalence when data are censored: a randomized trial of treatments for B non-Hodgkin lymphoma. Statist. Med. 12, 1353–1364.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman & Hall, New York.

Dunnett, C.W., Gent, M., 1977. Significance testing to establish equivalence between treatments with special reference to data in the form of 2×2 tables. Biometrics 33, 593–602.

Fleming, T.R., 1990. Evaluation of active control trials in AIDS. J. Acq. Immun. Def. Synd. 3 (Suppl. 2), 582–587.

Kaplan, E.L., Meier, P., 1958. Non-parametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457–481.

Nam, J.-M., 1987. A simple approximation for calculating sample sizes for detecting linear trend in proportions. Biometrics 43, 701–705.

Nam, J.-M., 1997. Establishing equivalence of two treatments and sample size requirements in matched-pairs design. Biometrics 53, 1422–1430.

Patte, C., Phillip, T., Rodary, C., Zucker, J.M., Behrendt, H., Gentet, J.C., Lamagnére, J.P., Otten, J., Dufillot, D., Pein, F., Caillou, B., Lemerle, J., 1991. High survival rate in advanced-stage B-cell lymphomas and leukemia without CNS involvement with a short intensive polychemotherapy: results from the French Pediatric Oncology Society of a randomized trial of 216 children. Amer. Soc. Clin. Oncol. 9, 123–132.

Roebruck, P., Kühn, A., 1995. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. Statist. Med. 14, 1583–1594.

Wellek, S., 1993. A log-rank test for equivalence of two survivor functions. Biometrics 49, 877–881.